IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Real-Time Voice Cloning

Raghav Joshi
Artificial Intelligence & Data Science
SIES Graduate School Of Technology
Navi Mumbai, India

Esakkimuthu Konar
Artificial Intelligence& Data Science
SIES Graduate School Of Technology
Navi Mumbai, India

Mayakrishna Yadav
Artificial Intelligence & Data Science
SIES Graduate School Of Technology
Navi Mumbai, India

Aditya Pai
Artificial Intelligence & Data Science
SIES Graduate School Of Technology
Navi Mumbai, India

Dr. Neethu Anna Sabu
Artificial Intelligence & Data Science
SIES Graduate School Of Technology
Navi Mumbai, India

Abstract: Replication of Voice using Real-Time Voice Cloning, which uses cutting-edge technology that enables the generation of a synthetic voice closely resembling the voice of the target speaker. This process involves two main stages: training the model on a large dataset of the target voice and utilizing that model to generate real-time speech. Contrary to traditional voice synthesis methods, which frequently require extensive computational resources and time for voice synthesis, real-time voice cloning aims to minimize latency and computational cost, making it feasible for practical applications such as voice assistants and accessibility tools. This paper presents an approach for real-time voice cloning, leveraging advanced techniques such as Generative Adversarial Networks, Wavenets, and Autoencoders. By optimizing both the training and synthesis stages, our approach achieves a superior synthesis of the target speaker's voice with minimum delay. We estimate the model's performance through various parameters such as intelligibility, naturalness, and speaker similarity. The results show that our methods offer a scalable solution for superior real-time Voice Cloning with the potential for diverse applications in both personal and professional domains. Applications range from personalized digital assistants to dubbing and entertainment, but ethical examination must be taken into account due to concerns about potential misuse in generating fake videos.

Index Terms - RTVC, Wavenets, Autoencoders, GANs, Vocoder.

I. INTRODUCTION

The rapid advancements in Machine Learning and speech synthesis have brought voice cloning technology to the forefront of Artificial Intelligence research. Voice Cloning refers to the process of replicating a person's unique voice characteristics, enabling the generation of natural-sounding speech that closely monitors the target speaker's voice traits such as pitch, tones, quality, accents, etc[4]. Historically, voice synthesis systems have required substantial computational resources and time-consuming processes, making them impractical for real-time applications[1]. However, recent breakthroughs in Deep Learning, particularly with the use of Generative Adversarial Networks, Recurrent Neural Networks (RNNs), and Autoencoders, have significantly reduced the time and resources for generating superior speech[6][7]. Real-time voice cloning aims to overcome the

limitations of traditional voice synthesis by enabling the immediate replication of a speaker's voice during interaction. This capability has far-reaching implications across various domains, such as personalized virtual assistants, accessibility tools for individuals with speech impairments, content creation, gaming, and even the entertainment industry[9]. By allowing dynamic and seamless voice generation with minimal delay, Real-Time Voice Cloning has the potential to transform human-computer interaction and open new frontiers for customized communication technologies[10]. This approach combines the element of speech synthesis to generate realistic and personal voices in real-time. The process involves three key stages: Encoder, Vocoder, and Decoder. The speaker encoder processes an audio sample of the target voice, extracting its unique features and generating a speaker embedding: a mathematical representation of the voice's characteristics, such as speech, tone, and speaking style. This embedding is then fed into the synthesizer, which then converts textual input to speech that closely relates to the voice of the target speaker. Finally, the vocoder converts the synthesized speech from a spectrogram format into a natural-sounding audio waveform[12].

Popular Python Libraries and frameworks such as Pytorch, Vocoder, and Speaker Encoder provide pre-trained models and essential tools to streamline the development process. Voice Cloning has a wide range of applications, from personalized virtual assistants to entertainment and accessibility tools. However, it also raises ethical concerns regarding privacy and misuse, requiring careful resolution. There has been a significant interest in end-to-end training of text-to-speech(TTS) models, which are trained directly from text-audio pairs without depending on intermediate representations. Tacotron uses WaveNet as a vocoder to invert spectrograms generated by an encoder-decoder architecture with attention, obtaining the naturalness of approaching human speech by combining Tacotron's prosody with WaveNet's audio quality[4]. This work aims to build a TTS system that can generate natural speech for a variety of speakers in a data-efficient manner. We specifically address a zero-shot learning setting, where a few seconds of untranscribed reference audio from a target speaker is used to synthesize new speech in that speaker's voice without updating any model parameters. Such systems have accessibility applications, such as restoring the ability to communicate naturally to users who have lost their voice and cannot provide many new training examples[11].

II. LITERATURE SURVEY

Real-Time Voice Cloning builds upon traditional speech synthesis and text-to-speech Systems(TTS). Early methods relied on concatenative and statistical parametric speech synthesis(SPSS), which required extensive datasets for training. Hidden Markov Models (HMMs) and Gaussian Mixture Models(GMMs) were commonly used but lacked adaptability.

A. Transfer Learning In NLP

Transfer Learning in NLP refers to the technique of leveraging pre-trained models on large datasets and fine-tuning them on specific tasks. Instead of training models from scratch, transfer learning enables knowledge reuse, reducing the need for extensively labeled data and improving performance. By adapting learned representations to new tasks with minimal additional training, transfer learning has revolutionized NLP, making it more efficient and accessible[3].

B.Synthesizing Informative Speech With Style and Emotion

Synthesizing speech with style and emotion involves generating speech that is not only clear and accurate but also engaging and expressive. By integrating elements such as pitch, variation, rhythm, and stress, these systems can enhance listener engagement. Techniques such as prosody transfer and effective computing enable speech synthesis models to adapt to different speaking styles, making them suitable for applications like virtual assistants and customer service interactions[4].

C. Transformers in NLP

Transformers have revolutionized Natural Language Processing(NLP) by introducing a self-attention mechanism that enables models to process entire text sequences in parallel. Transformers can capture long-range dependencies efficiently, making them highly effective for tasks like machine translation, text generation, and sentiment analysis[10].

D. Tacotron: Towards End-to-End Speech Synthesis

Tacotron is an end-to-end speech synthesis model developed by Google that generates natural-sounding speech directly from text. Unlike traditional text-to-speech (TTS) systems that require complex pipelines,

Tacotron simplifies speech synthesis by using a sequence-to-sequence architecture with attention mechanisms. It maps character sequences to spectrograms, which are then converted into waveform audio using a Vocoder like WaveNet. The model improves speech naturalness by capturing prosody, intonation, and rhythm effectively. Tacotron's advancements laid the foundation for more refined models,
such as Tacotron 2, which further enhanced speech quality and realism[2].

III. PROPOSED SYSTEM

The proposed system for real-time voice cloning aims to enhance the quality, adaptability, and efficiency of voice synthesis by integrating advanced neural network architectures and techniques. This system will consist of three main components: a speaker encoder, a text-to-mel encoder, and a vocoder. Here's an overview of each component:

A. Speaker Encoder

The speaker encoder will be responsible for extracting distinctive characteristics from a limited number of audio samples from the target speaker. This component will utilize techniques such as few-shot learning to enable the system to clone the voice with minimal data. By implementing a robust embedding mechanism, the encoder will ensure that the synthesized voice closely resembles the target speaker, allowing for flexibility in applications where only a few samples are available.

B. Synthesizer

The synthesizer is typically based on sequence-to-sequence models such as Tacotron2 or FastSpeech that convert the text into Mel- Spectrogram. The input to the Synthesizer is a raw text or input speech, which is converted into phonemes, generally used to improve the pronunciation. These phonemes are encoded using the embedding layer like the word embedding in NLP. These embeddings are further passed in the Recurrent Network, which is LSTM. The model then aligns the input phonemes with the output of the spectrogram. This is followed by the Decoder taking the aligned features and generating the spectrogram. Thus Synthesizer outputs the characteristics of the voice[5].

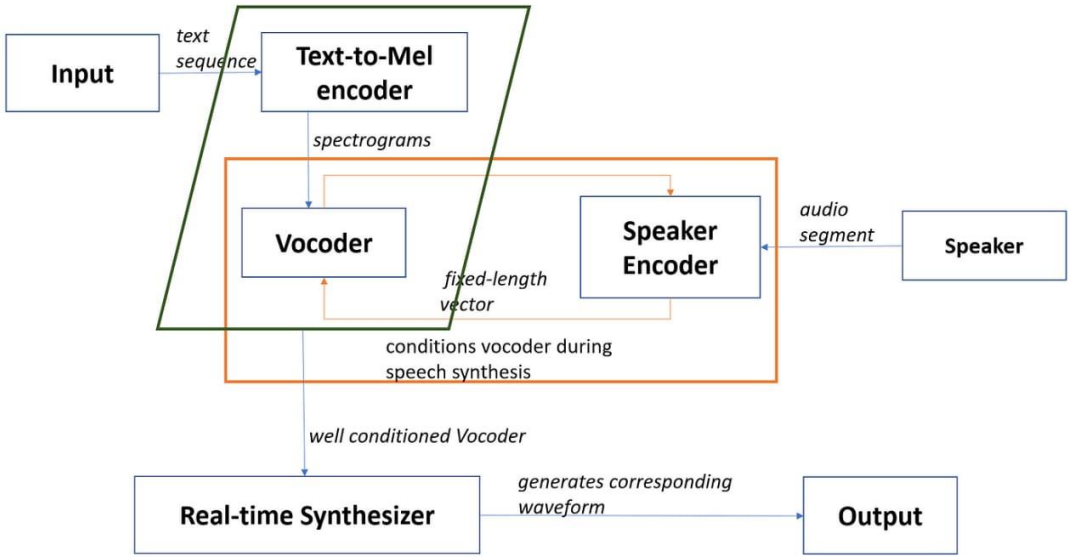

Fig 1. Proposed System

C. Neural Vocoder

The Neural Vocoder converts the Mel Spectrogram into Waveforms by modeling the complex details of the human speech, such as pitch, quality, and articulation. The Neural Vocoder uses Deep Learning to predict the waveform from the spectrogram and produces a high-fidelity audio waveform that sounds natural. These Neural Vocoders are powerful since they learn complex speech patterns better than traditional vocoders. These

Neural Vocoders enable the production of real-time speech synthesis, voice cloning, Text-to-Speech, and Speech Enhancement[1].

IV. METHODOLOGY

The real-time voice cloning system comprises several key algorithms that work together to achieve high-quality, natural-sounding voice synthesis. The following are the mechanisms involved in each part of the system.

A. Speaker Encoder

A Speaker Encoder is a Deep Learning model that extracts voice embeddings from a short voice sample. These embeddings capture the unique vocal characteristics of voice such as pitch, tone quality, etc.

Workflow Of Encoders:

• Audio Processing
– Convert raw speech into Mel- Spectrogram.
– Normalize the spectrogram to reduce noise sensitivity.

• Embedding Extraction
– Pass the Mel-Spectrogram through a Deep Neural Network.
– The Network learns a compact fixed-dimensional embedding.

• Loss Function
– Generalized End-to-End (GE2E):- Forces embeddings from same speakers to be closer while different speakers remain separated.
– Triplet Loss:- Encourages embeddings of same speakers to be closer than different speakers.

• Using the Speaker Embedding
– The extracted speaker embedding is sent to the Synthesizer.
– The Synthesizer conditions the text-to-speech model to generate speech in a cloned voice.

• Generalized End-to-End (GE2E) Loss Formula

$$L = \sum_{i=1}^{N} \sum_{j=1}^{M} (d(s_{i,j}, c_i) - d(s_{i,j}, c_k))$$

- $s_{i,j}$ is the embedding of speaker i and embedding j.
- d is the cosine distance
- c_i is the centroid of speaker i
- c_k is the centroid of speaker k

B. Synthesizer

A synthesizer in Real-Time Voice Cloning is accountable for converting text and voice embeddings into a Mel-Spectrogram. The Synthesizers ensure that the generated speech matches both linguistic content and speakers' voice characteristics.

Workflow of Synthesizer:
• Text Processing.
– Converts input text into a phoneme sequence.
– The text sequence is embedded into a continuous vector representation.

• Sequence-to-Sequence Mapping
– The model uses a Deep Learning architecture to convert the text embedding into a Mel-Spectrogram.

Speaker Adaptation
– The Synthesizer takes speaker embedding from the speaker encoder.
– Conditions the Text-to-Speech model to mimic the speaker's voice.

• Mel Spectrogram Generator
– The decoder outputs a Mel-Spectrogram, capturing the speech's pitch, duration, and Energy[8].

C. Neural Vocoder
A Neural Vocoder converts a Mel-Spectrogram into a high-quality audio waveform. These are used in place of traditional Vocoders as they are faster and produce natural-sounding speech with realistic prosody. Working Of Neural Vocoders:

• Input Feature Processing.
– The Mel-Spectrogram is unsampled to match the resolution of the waveform, which ensures that the network gets enough time domain information.

• Loss Function for Neural Vocoders

$$L_G = L_{adv} + \lambda L_{fm} + \mu L_{mel}$$

Where :

- L_G is the total loss function for training the Neural Vocoder or speech synthesis model.
- L_{adv} is the adversarial loss to make waveforms to sound real.
- L_{fm} is the feature matching loss for stability.
- L_{mel} is the Mel-Spectrogram reconstruction loss
- λ and μ are the weighting factors that balance the contributions of different loss functions
- λ and μ control the feature matching loss L_{fm} and the Mel-Spectrogram loss L_{mel}.
- If λ is high, the model focuses more on the feature matching loss and produces smoother speech. If λ is low, the model relies more on adversarial loss, which can lead to more realistic but potentially less stable results.

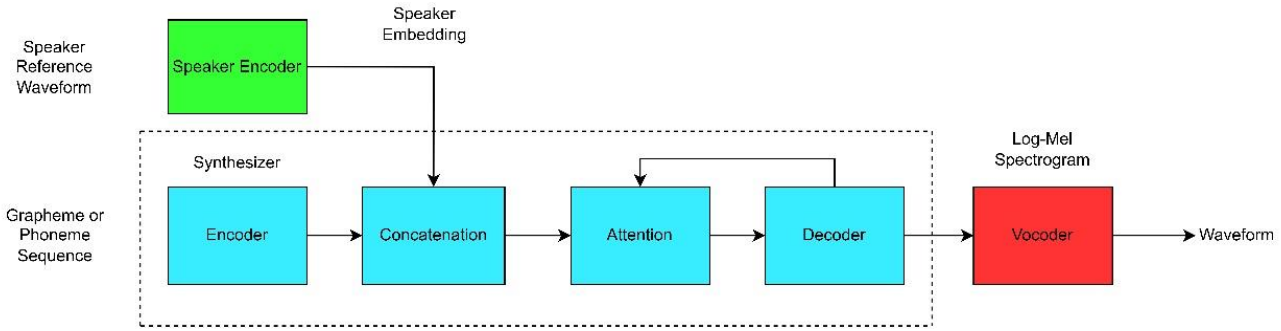

Fig 2. Methodology Of Real-Time Voice Cloning

V. RESULTS AND DISCUSSION

- In Results, we performed various experiments on the Cloning System and found that the Cloning system we developed required five seconds to produce high-quality audio. The model developed by us uses WaveRNN, which has less computational cost as compared to other models that use Tacotron and Wavenet, which have high computational costs. The separation between the Encoder, Synthesizer, and Vocoder components allows potential and flexible enhancements in each module independently. Also, the system lacks security as it enables the instantaneous replication of someone's voice.

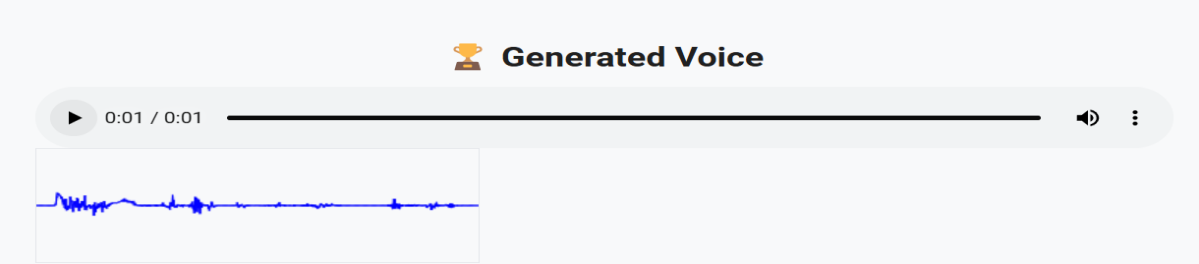

Fig 3. Waveform Of Generated Voice

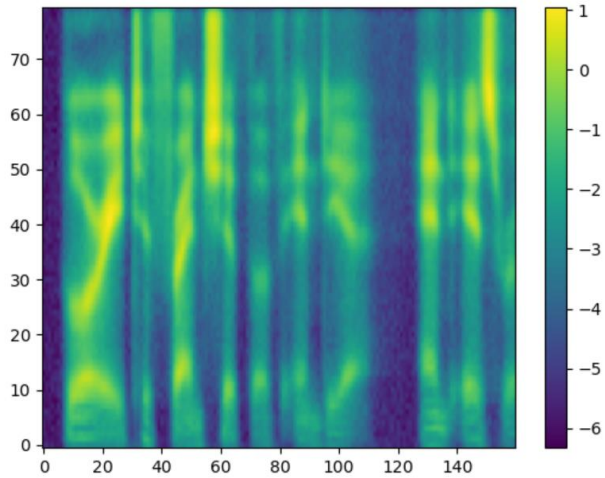

Fig 4. Mel-Spectrogram generated from Synthesizer

- The result of this study highlights the effectiveness of the combination of different Encoder, Synthesizer, and Vocoder models to minimize the loss and improve the various characteristics of voice such as Pitch, Articulation, Quality, etc. The analysis of the components such as the Encoder, Synthesizer, and Vocoder and the comparison of different Neural Networks used in these components help to improve the characteristics of the voice produced. The main aim is to compare various Neural Networks used in the Encoder and combine them with different Neural Networks of Vocoder and an efficient Synthesizer to improve the voice produced after cloning the speech or text that is required to be cloned.

  • **Speaker Encoder**:
  A Generalized End-to-End Speaker is used as Speaker Encoder, which uses an LSTM network to extract time-dependent speaker features and generate speech in the target voice. GE2E speaker encoder is faster and has higher accuracy than other encoders, and is used in Real-Time Voice Cloning.

  • **Synthesizer:**
  Tacotron2 is chosen as a synthesizer that uses Recurrent Neural Networks with Attention Mechanisms and aligns input phonemes with output Mel-Spectrogram frames. It is slower but has superior quality.

• **Neural Vocoder:**
We use Wave-RNN for Neural Vocoder, which uses a Recurrent Neural Network and is optimized for low-powered devices. Alternatively, we can use Hifi-GAN, which uses GANs where the generator converts spectrograms to waveforms and multiple Discriminators evaluate the realism of the generated audio. Therefore, it is very fast.

VII. CONCLUSION

In conclusion, we present a neural network-based system for multi-speaker Text-To-Speech synthesis. The system combines an independently trained speaker encoder network with a sequence-to-sequence Text-To-Speech synthesis network and neural vocoder based on Tacotron2. By leveraging the knowledge learned by the discriminative speaker encoder, the synthesizer is able to generate high-quality speech not only for speakers seen during training but also for speakers never seen before. Through evaluations based on a speaker verification system as well as subjective listening tests, we demonstrated that the synthesized speech is reasonably similar to real speech from the target speakers, even on unseen speakers. We ran experiments to analyse the impact of the amount of data used to train the different components and found that given sufficient speaker diversity in the synthesizer training set, speaker transfer quality could be significantly improved by increasing the amount of speaker encoder training data. Finally, we demonstrate that the model can generate realistic speech from fictitious speakers that are dissimilar from the training set, implying that the model has learned to utilize a realistic representation of the space of speaker variation.
The proposed model does not attain human-level naturalness, despite the use of a WaveNet vocoder (along with its very high inference cost), in contrast to the single-speaker results.

REFERENCES

[1] Oord, A. V. D., et al. (2016). ”WaveNet: A Generative Model for RawAudio.” arXiv preprint arXiv:1609.03499.
[2] Wang, Y., et al. (2017). ”Tacotron: Towards End-to-End Speech Synthesis.” Interspeech 2017.
[3] Jia, Y., Yu, Z., Wu, Y. (2018). ”Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis.” Interspeech 2018.
[4] Shen, J., et al. (2018). ”Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions.” 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
[5] Kong, J., Zhang, Y. (2019). ”FastSpeech: Fast, Robust and Controllable Text to Speech.” Advances in Neural Information Processing Systems.
[6] Nielsen, J., et al. (2019). ”Real-Time Speech Synthesis Using Deep Learning.” International Journal of Computer Applications.
[7] Huang, Y., et al. (2020). ”Generative Adversarial Networks for Voice Conversion.” IEEE Transactions on Audio, Speech, and Language Processing.
[8] Li, N., et al. (2020). ”Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram.” Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
[9] Chen, J., et al. (2020). ”Real-Time Voice Cloning.” arXiv preprintarXiv:1806.04558.
[10] Desai, N., et al. (2020). ”Few-Shot Voice Cloning with Attention.”ICASSP 2020.
[11] Zhang, Y., et al. (2020). ”Unsupervised Voice Cloning for Zero-Shot Text-to-Speech.” Interspeech 2020.
[12] Dufour, C., et al. (2020). ”Few-Shot Voice Cloning with a SpeakerAware Text-to- Speech Model.” Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).